ACS SyntheticBiology

Queen Mary University of London | Library Services

Subscriber access provided by Queen Mary, University of London

Article

Programmed protein self-assembly driven by genetically encoded intein-mediated native chemical ligation

Joseph A Harvey, Laura S. Itzhaki, and Ewan Main

ACS Synth. Biol., **Just Accepted Manuscript** • DOI: 10.1021/acssynbio.7b00447 • Publication Date (Web): 23 Feb 2018

Downloaded from http://pubs.acs.org on March 2, 2018

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.

ACS Publications

ACS Synthetic Biology is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Published by American Chemical Society. Copyright © American Chemical Society. However, no copyright claim is made to original U.S. Government works, or works produced by employees of any Commonwealth realm Crown government in the course of their duties.

Programmed protein self-assembly driven by genetically encoded intein-mediated native chemical ligation.

Joseph A. Harvey[1], Laura S. Itzhaki[2] and Ewan R.G. Main[1]*

[1]School of Biological and Chemical Sciences
Queen Mary, University of London,
Mile End Road
London E1 4NS, U.K

[2]Department of Pharmacology
University of Cambridge
Tennis Court Road
Cambridge CB2 1PD, U.K.

*To whom correspondence should be addressed, email: e.main@qmul.ac.uk

Abstract

Harnessing and controlling self-assembly is an important step in developing proteins as novel biomaterials. With this goal, here we report the design of a general genetically programmed system that covalently concatenates multiple distinct protein domains into specific assembled arrays. It is driven by iterative intein-mediated Native Chemical Ligation (NCL) under mild native conditions. The system uses a series of initially inert recombinant protein fusions that sandwich the protein modules to be ligated between one of a number of different affinity tags and an intein protein domain. Orthogonal activation at opposite termini of compatible protein fusions, via protease and intein cleavage, coupled with sequential mixing directs an irreversible and traceless stepwise assembly process. This gives total control over the composition and arrangement of component proteins within the final product, enabled the limits of the system - reaction efficiency and yield - to be investigated and led to the production of “functional” assemblies.

Keywords: directed protein assembly / native chemical ligation / traceless protein conjugation / expressed intein ligation / protein design / nanostructures

Self-assembly systems are powerful tools found throughout Nature for building highly ordered and functional structures from simple starting materials (some examples include collagen, tubulin, laminin, and actin). As such, manipulation and mimicry of these systems is an excellent step in the development of novel man-made functional and responsive biomaterials for biotechnological applications [1-6]. Proteins are arguably one of the best building blocks for such designed self-assembly systems. This stems from their diverse structures and functions, coupled with their ability to be functionalised and easily produced by recombinant methods [7-8]. However, controlling the assembly of protein units on cue into ordered 1-D, 2-D and 3-D structures with specific directionality and precise control of spatial integration is an extremely difficult task. Successful controlled assembly systems in Nature and design thus far have tended to use both non-covalent and covalent mechanisms that exhibit the following characteristics: (i) a driving force to encourage specific interactions between proteins, such as protein-protein interfacing or orthogonal covalent ligation, (ii) a mechanism to order the assembly, for example symmetry or repetitive docking and (iii) ideally a control mechanism to initiate or terminate the assembly reaction.

Although there is a limited number of synthetic bioconjugation chemistries that can be used to control protein and peptide assembly, their scale-up costs, labelling limitations and orthogonality minimise their effectiveness in assembling larger protein arrays [9]. Genetically programmed assembly can circumvent a number of these challenges by recombinant expression of protein modules. Successful engineering of such systems, and smaller peptides via solid-state synthesis, has mainly focused on assembly through one of the following: compatible orthogonal non-covalent interfaces such as dimer-trimer protein fusions, the use of orthogonal tags (e.g. coiled coils), protein-ligand binding or via computational interface design [3, 5, 10-13]. The most commonly used means of genetically encoded covalent control has been through the use of disulphide linkages and thus redox chemistry [3, 14-17].

Truly irreversible genetically programmed covalent bonding would widen the potential application as well as the molecular size of assemblies that can be

accessed. As, for example, these approaches provide a route to linking proteins into single polypeptide chains that circumvents both protein misfolding/protein synthesis errors associated with recombinantly producing extremely large polypeptides [18-19] and technically challenging construction/recombinant production of repetitive and/or cytotoxic sequences [20-21]. Thus they enable the construction of protein assemblies with structures and functions as diverse as linking of multiple enzymes into nano-reactors [22], site specific modification of antibodies [23] and synthetic vaccine construction [24-26].

At present irreversible genetically programmed covalent bonding has been limited to a few specific enzymatically controlled cross-linking reactions [16, 27-34]. The most relevant for producing larger arrays have used: intein-mediated native chemical ligation (NCL) which produces a peptide bond between N- and C-termini of proteins [27-28] and de-constructed adhesion domains from various bacteria which "self-catalyse" isopeptide bond formation between an aspartic acid and lysine side chains [16, 30, 32]. Although a designed sequential isopeptide bond-forming system has successfully produced large linked protein assemblies and has found exciting multiple uses [16, 24-26, 30, 32], the method does leave behind a whole protein domain at the site of each ligation. In contrast, we recently designed an intein fusion protein system that polymerises protein modules tracelessly into single protein chains without leaving any unwanted polypeptide behind (Figure 1) [19]. Our one-pot synthesis system genetically encodes triggerable directional assembly via orthogonal chemistries at the termini of the module to be polymerised [an N-terminal cysteine revealed by protease cleavage and a C-terminal thioester produced from the excision of a Mxe GyrA intein (Mycobacterium xenopi DNA gyrase A) (Figure 1A)]. Here the Mxe GyrA intein acts both as a "protecting group" (inert until catalysed to self-excise with added reducing agent) as well as one half of the activating chemistry. This is in contrast to "natural intein ligation" where the intein is usually found in the middle of a gene and then post-translationally self-excises itself whilst ligating the flanking polypeptides regions together. Although our one-pot synthesis was successful in producing single protein chains that extended up to microns in length,

it was not possible to position specific units at a predefined position or produce assemblies of defined lengths.

Here we present a recombinant protein system that uses intein-mediated NCL to assemble natively folded proteins in mild native conditions in a traceless irreversible stepwise process with precise control. The system is driven via Mxe GyrA intein and is conceptually similar to stepwise peptide synthesis. It requires a minimum of two different fusion proteins that possess the protein units to be assembled inserted between differing flanking affinity protein tags. The affinity tags allow the masking of the reactive groups, chemical attachment of the reactants and/or products to a solid support (either a SPR chip or agarose beads) and easy separation of the product from unreacted starting components. Used in a stepwise process, the fusions gave total control of the composition and arrangement of component proteins within the final assembled product and permitted the limits of the system to be investigated.

Results.

System Design: Our two-component system is based on recombinantly expressing the protein modules to be assembled in fusion constructs termed “anchor” and “linkers” (Figure 2). At the N-terminus each fusion construct contains an affinity tag that protects one half of the orthogonal NCL chemistry (a reactive cysteine residue) and enables easy purification (the anchor contains a GST tag whilst linkers contain a polyhistidine tag). At the C-terminus the two fusions differ more fundamentally. The linkers contains a C-terminal Mxe GyrA intein to produce the complimentary half of the orthogonal NCL chemistry (a thioester) and a chitin binding domain affinity tag (CBD) for purification. The Mxe Gyr A Intein was chosen as it has high expression levels in bacteria and can be purified in denaturing conditions and then easily refolded back into its native state. It also tolerates the presence of low levels of detergents and denaturants during thiol-induced cleavage. In comparison, the anchor contains no reactive C-terminal domains and instead contains a C-terminal StrepTactin tag for resin attachment if solid state synthesis is required. Cloning,

expressing and purifying different proteins in these constructs produces natively folded and initially inert fusions. Selective activation at opposing termini coupled with irreversible stepwise assembly, driven by NCL, enables directional construction of poly-proteins with specific compositions and spatial arrangements. The fabrication process begins with the anchor fusion and proceeds from its N-terminus allowing the build-up of the protein nanostructures as follows:

Fabrication via Iterative Stepwise Additions

Activation: The anchor fusion is N-terminally activated through protease cleavage of its affinity tag. The deprotection produces an N-terminal reactive cysteine (Figure 2 & 4A). Separately, the linker fusion is C-terminally activated via intein self-cleavage induced by addition of the reducing agent MESNA. This produces a reaction-ready C-terminal thioester (Figure 2).

Stepwise Addition: The orthogonal C-terminal thioester activated linker is incubated with N-terminal cysteine activated anchor resulting in a spontaneous irreversible NCL reaction (Figure 2). The reaction creates a "dimeric" protein with a traceless new peptide bond between the anchor at the C-terminus and the linker at the N-terminus. The ligated "dimeric" product has an affinity tag combination distinct from either of the two reactants, allowing it to be easily isolated.

Further Stepwise Additions: During the previous ligation the linker's N-terminal cysteine remains protected via its affinity tag (Figure 2). This protected cysteine can be liberated as previously via protease cleavage making it reaction-ready. A ligation step can then be repeated by mixing with a new C-terminally activated linker fusion (containing whichever protein is required). Again the product contains a distinct affinity tag combination for easy purification and can be activated at its N-terminal for further iterative additions to occur. Thus, through a series of sequential deprotection and NCL reactions the system can iteratively construct a novel protein nanostructure. Moreover, as the assembly can be isolated at each step it will consist of a single species of exact size and composition.

Protein Ligation Optimisation

High ligation yields are required due to sequential iterative nature of the assembly process. To determine the optimum native conditions for reaction yield, and thus the limits of the system, we chose to ligate our model proteins of choice: CTPRs (Consensus Tetratricopeptide repeat proteins). Initially identical CTPR3ΔS modules were cloned into the anchor and linker constructs. The CTPR3ΔS modules contained 3 consensus repeats that lacked a binding interface (termed a "spacer"). The spacer CTPR3ΔS protein has been shown to be highly thermodynamically stable and highly resistant to aggregation [28, 35]. Both the anchor and linker were expressed in high yields with the spacer CTPRs (10-20 mgs per litre), were easily purified to > 95 % and activated with yields of > 90 % (SI Figure 1, Materials and Methods). Ligation optimisation was carried out by varying the following parameters: using different proteases to activate the N-terminal cysteine, changing the protein concentrations and ratios, pH, salt, temperature, thiol catalysts, ligation time and time between activation and initiating ligation. The results (Figure 3) show that the most important elements to high yielding ligations are:

(i) Rapid, specific production and use of the N-terminal cysteine – Specific activation coupled with the time and conditions between protease cleavage and ligation of the cysteine were critical. Non-specific protease cleavage or disulphide formation reduced the available react-able cysteines and the longer the correctly cleaved/activated cysteine was left at room temperature before ligation, the lower the reaction yield (Figure 3A). Thus the highest yielding ligations were achieved when protein was activated quickly in a highly reduced environment and stored at -80 °C. Of the protease's trialled (TEV, Thrombin and Factor Xa), both Thrombin and Factor Xa slowed significantly in the presence of reducing agent and, in the case of Factor Xa, produced a significant percentage of promiscuous cleavage (SI Figure 2). Only TEV gave the specific and fast cleavage in the reducing environment required (SI Figure 3). Therefore, all our fusion designs only possess TEV cleavage sites and were performed with TCEP to keep a highly reduced environment.

(ii) Optimum temperature, protein concentration & pH – After the importance of the N-terminal cysteine, temperature, pH and protein concentration were found to be the most important factors. The optimum reaction temperature and pH were found to be 30 °C and pH 8.5, higher temperature/pH decreased yield through non-specific aggregation and degradation (Figure 3C & D). In a similar manner, synthetic peptide to peptide NCL studies have shown that high reaction yields are achieved the use of a high concentration and large molar excesses of attacking peptide reactant (usually > 1 mM and in a 10:1 ratio [21]). This is generally carried out in the presence of strong denaturants to prevent aggregation [for example, 6 M Guanidine Hydrochloride (GuHCl)]. Clearly such harsh conditions are undesirable for protein self-assembly; as this would cause protein denaturation and thus require added refolding complications. The high stability and resistance to aggregation (can be concentrated to the low mM range) of the CTPR3ΔS spacer module enabled the NCL reaction to be trialled over a range of protein concentrations and molar ratios under native conditions (Figure 3G, SI Fig 4). At higher concentrations the yields were significantly higher, and increasing the molar ratio of thioester to cysteine reactant similarly improves percentage yield. However, such high concentrations can only be achieved with the most "amenable" protein domains. Therefore protein concentrations in the µM range were chosen (100 µM spacer to 50 µM anchor) in a 2:1 ratio (thioester to cysteine).

(iii) Identity and buffer conditions of the C-terminal thioester – In contrast to the N-terminal cysteine, the MESNA produced C-terminal thioester remains highly active for at least 3 days, with significant loss of reactivity only after 7 days (when in an excess of MESNA) (Figure 3B). Peptide studies have shown the identity of the thioester can affect reaction rates, therefore the MESNA formed thioesters were exchanged through incubation with the known activator 4-carboxymethyl thiophenol (MPAA) and deactivator dithiothreitol (DTT) [36]. In line with these studies, ligation rates and initial yields increased with MPAA or were greatly reduced with DTT (Figure 3). However, increasing the time of ligation to 24 hours and using a higher concentration of MESNA produced similar yields to the 16 hour MPAA incubated reactions. Thus ligations were performed with less expensive MESNA formed

thioester for 24 hours in buffers containing > 100 mM MESNA and where all DTT was replaced with TCEP. Finally, most of the solubilising additives trialled had little effect, except for NaCl and glycerol which increased and decreased the reaction yield slightly, respectively (Figure 3H).

Sequential Ligations and nanostructure formation

From the optimisation studies above, final ligation conditions of 100 µM activated thioester linker to 50 µM activated cysteine anchor in 50 mM Tris-HCl, 500 mM NaCl, 200 mM MESNA, 10 mM TCEP, pH 8.5, 30 ˚C for 24 hours were used. An average final reaction yield of approximately 75 % ± 5.5 was achieved (SI Figure 4). This suggests that the highest feasible number of sequential ligations that could be performed would be 6 and produce a final yield of ≈ 18 %. However, although the yield for each NCL is high, it is still within the range that necessitates purifying the product from the reactants after each reaction step (avoiding cross reactivity in future ligation steps). The additional purification requirement results in good overall recovery of ligation product (75 to 50 %). A higher recovery was always obtained when new StrepTactin® resin (resin-bound tetrameric streptavidin) was used, with greater losses as the resin aged. Therefore, combined yield after each stepwise addition is ~50 %, suggesting that 3 to 4 sequential ligations joining 4 to 5 protein modules together is the realistic limit of the system.

To show such feasibility we therefore undertook a series of ligation reactions using the spacer CTPR3ΔS. Figure 4B shows a typical three sequential ligation sequence (the reactants and products of each ligation were verified by Mass Spectrometry – SI Figure 1 & 5). The yield for each individual NCL addition was consistent with previously obtained yields and after each ligation the product was successfully purified to > 90 % and re-activated to > 95 %. The final product, a semi-synthetic CTPR12ΔS protein consisting of 4 CTPR3ΔS subunits, was > 90 % pure and correctly folded (as monitored by Far UV Circular Dichroism – SI Fig. 5). Moreover, as a CTPR domain of between 18 and 20 motifs can be reliably expressed recombinantly, a protein fibril of at least 60 motifs or more, equating to ~50-70 nm in length, could be reliably produced.

Functionalising Ligated Nano-structures

Given the success of the bare-bone ligations, the next step was to introduce functionality to the nanostructures. This was achieved by placing new "binding" CTPR modules at specific spatial locations within the fabricated CTPR ensemble. The new module consisted of a CTPR6ΔS monomer constructed from a "spacer" CTPR3ΔS unit fused to a "binding" CTPR3ΔS unit (S.I Figure 1C). The "binding" CTPR3ΔS domain chosen was designed and developed by Regan and co-workers and was capable of associating with the pentapeptide amino acid sequence – MEEVD [37]. The CTPR6ΔS gene was cloned into the linker fusion and was recombinantly expressed at yields of 5 - 10 mgL^{-1}. The expressed fusion was then activated to a thioester monomer and purified to > 95 % (S.I Figure 1C). To form a "functional" nanostructure, i.e. one with binding capability, three stepwise NCL reactions were performed with the spacer-containing anchor and "binder" CTPR6ΔS-containing linkers. Reactions were performed as previously, except the ligation temperature was reduced to 25 °C to prevent aggregation (the "binder" construct was more prone to aggregation than the "spacer" constructs at higher concentrations). Figure 4C shows a 3 sequential reaction fabrication with the final CTPR21ΔS protein (≈ 92 kDa) confirmed by ESI Mass Spectrometry (SI Figure 5). As with the previous stepwise ligations, the product of each ligation was separated from its reactants (70-95 %) and reactivated to >95 %. When the ligation yields of these reactions were compared with the sequential spacer additions performed at 30 °C, they were found to be lower. This was expected and in line with yields of the "spacer" when ligated at 25 °C. The newly synthesized CTPR21ΔS protein produced is directly comparable to the CTPR18 protein that Regan and co-workers used to form smart hydrogels through mixing with a 4-armed PEG-peptide linker [4]. Thus, by ligating specific modules (in this case binding and spacer CTPRΔS domains) at predetermined spatial locations, our system can produce a semi-synthetic functional nanostructure.

Discussion.

Herein we have developed a general recombinant protein system that utilises genetically programmed intein-mediated NCL to irreversibly and sequentially assemble large user-defined protein architectures from smaller protein domain building blocks. Under mild conditions natively folded protein domains react to produce a traceless peptide-bond linked product (with a purity of > 90 %) with no intervening tag and no requirement for chemical modification. Thus no bioconjugations or refolding steps are required. Furthermore, at the relatively low protein concentrations used (100 μM C-terminal thioester to 50 μM N-terminal cysteine reactants), the system should to be widely accessible to numerous protein systems. With such reaction conditions each sequential ligation step was optimised to yield 75 % product, generating a realistic limit of the system of 3 to 4 sequential ligations joining 4 to 5 protein modules together. Proteins that are soluble up to higher concentrations would increase this yield even further and enable an expansion to a greater number of sequential steps. Moreover, as the anchor construct possesses a C-terminal StrepTactin tag the system could be attached to resin if solid state synthesis via, for example, Biacore™ is required.

Using the assembly system we have shown that modules made from 3 or 6 consensus TPR motifs can be natively assembled into functional proteins of up to 21 CTPR motifs (≈ 92 kDa). In particular, this strategy allowed us to place binding modules at specified locations to enable gelation with a four-armed PEG-peptide linker. Thus, our system provides: (i) a novel semi-synthetic shortcut to the production of giant repetitive proteins that are challenging to produce via molecular biology and (ii) has the potential to create customised protein-based biomaterials comprising multiple functionalisable protein elements that can be specifically orientated within the final protein architecture. One future avenue would be to create two- and three-dimensional lattices or cages through addition of oligomeric domains into the anchor and linker constructs. Other uses could include the semi-synthesis of cytotoxic proteins from multiple non-toxic protein units or for the synthesis of protein multimers that are too large to be produced recombinantly as

single polypeptide chain (≈ 100 kDa *E.coli*). Furthermore, this semi-synthetic approach would enable additional modifications to be incorporated that are difficult to achieve via recombinant techniques. Such modifications could include the insertion of a monomer unit containing a non-natural amino acids (for example, azido-amino-acids used in Click chemistry) which could enable attachment of other components to the protein scaffold.

Methods.

Construction of Fusion Protein Genes & Vectors: A detailed description of the construction and cloning of the fusion protein genes used are described in the Supplementary Information (SI).

Protein Production and Purification: All fusion proteins were obtained by recombinant *E. coli* expression and purification by standard affinity chromatography techniques used previously [28] or outlined in Supplementary Information (SI).

Activation for Native Chemical Ligation: To make each fusion protein reaction-ready, they were activated at either the N- or C-termini via: (i) cysteine liberation or (ii) thioester formation, respectively:

(i) N-terminal Activation: The N-terminal cysteine was liberated by removing the protective affinity tag by protease cleavage. Three proteases were trialled: Tobacco Etch Virus (TEV), Factor Xa and Thrombin. Both TEV and Thrombin required modified cleavage sequences to liberate the N-terminal cysteine. The TEV protease cleavage was mutated to ENLYFQ↓C, from the commonly used site ENLYFQ↓G and the Thrombin protease sequence was mutated from LVPR↓(G/S) to LVPR↓C. Factor Xa cleaves directly after the 4 amino acid sequence IEGR, therefore the native cleavage site was utilised. 100 µM protein fusions were cleaved at 25 °C in 50 mM Tris-HCl pH 8, 150 mM NaCl, for 16 hours (10 mM TCEP was added to the reaction mixture for TEV cleavage and 0.1 mM TCEP for both Thrombin and Factor Xa). Post-cleavage affinity chromatography, using either Ni-NTA or GST resin depending on the N-terminal tag, facilitated the easy removal of both the cleaved affinity tag and un-cleaved fusion protein. After purification, the activated proteins were flash frozen and stored at -80 °C.

(ii) C-terminal Activation: The C-terminal thioester was formed by Sodium 2-mercaptoethanesulfonate (MESNA) triggered cleavage of the C-terminal *Mxe* GyrA intein fusion. Proteins were either bound to chitin resin and cleaved via incubation at 25 °C for 16 hours with cleavage buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 10 % w/v MESNA) or cleaved in the cleavage buffer at 100 µM concentration and then

purified via chitin resin. In both cases the C-terminal thioester-containing cleaved protein does not bind the chitin resin and is eluted, whereas the excised intein and un-cleaved fusion protein remains attached to the chitin beads via its chitin binding domain affinity tag. The formation of the C-terminal thioester was estimated to be 80 % or 95 % efficient depending on whether cleaved on or off the chitin resin, respectively. The cleaved thioesters were > 90 % pure after elution from the chitin resin (by SDS PAGE analysis S.I Figure 1). After purification, the activated proteins were flash frozen and stored at -80 °C.

Protein Ligation Reactions: Protein ligations were carried out in a ligation buffer of 50 mM Tris-HCl, 150 mM NaCl, 5 % (w/v) MESNA and 10 mM TCEP. Differing protein concentrations and ratios, pH, salt, temperature, thiol catalysts, ligation time and time between activation and initiating ligation were then investigated. The reactions were left to proceed under mild agitation. A small amount of urea (0.1-0.5 M) was added to the buffer to prevent any trace amounts of TEV protease causing any premature N-terminal activation of reactants.

Reaction Yields: All purification steps, cleavage and ligation reactions were monitored and confirmed by SDS-PAGE electrophoresis and either MALDI or Electrospray Mass Spectrometry (SI Materials & Methods). An Odyssey LI-COR in 800 nm imaging channel was used to quantify values for reaction yields of purification, cleavage and ligation reactions from the SDS-PAGE gels. Integrated intensity values (I) corresponding to each protein band were thereby obtained, and equation 1 was used to obtain the percentage of N-cysteinyl-containing protein successfully ligated to its thioester containing NCL reaction partner:

$$\%\ \text{Yield} = \left\{\left(\frac{\text{MolecWt C}}{\text{MolecWt L}}\ .\ \text{IL}\right)/\left[\left(\frac{\text{MolecWt C}}{\text{MolecWt L}}\ .\ \text{IL}\right)+\text{IC}\right]\right\}.100 \qquad (1)$$

where MolecWt C is the molecular weight of N-cysteinyl reactant, MolecWt L is the molecular weight of NCL product, IC is the Integrated Intensity of N-cysteinyl reactant and IL is the Integrated Intensity of NCL product. Equation 1 assumes that

the binding of Coomassie stain (and, therefore, the intensity) is linearly related to the molecular weight of each NCL reactant protein. This is a reasonable assumption given that all reactants are consensus tetratricopeptide repeat containing proteins (CTPR).

Accuracy and reproducibility of the ligation reactions: The accuracy and reproducibility of the ligation reactions was determined in triplicate, with 8 experiments performed 5 times. Reproducibility of the ligation reaction between protein preparations was also investigated by comparing optimised reaction condition yields between 4 differing protein preparations. In all cases the standard deviation for each experiment did not exceed 6.5 % of the yields reported and generally produced a standard deviation of between 1 and 2.5 % (SI Figure 4E). When comparing ligation yields between protein preparations, the standard deviation produced was 3.7 % (SI Figure 4E). Thus, the experimental yield analysis is robust, with the variation between protein preparations within the error of the experiment and analysis.

Supporting Information

Supplementary Materials and Methods: construction of fusion protein genes and vectors, expression and purification of fusion proteins, protein storage, confirmation of products and reaction yields, N-terminal activation of differing proteases and their efficiency.

Supplementary Figures: SDS PAGE gel / mass spectrometry analyses of the purification and activation of various anchor and linker fusions with differing proteases, SDS PAGE analyses of one-pot polymerisation reactions activated with various proteases, SDS PAGE analyses of the repeatability of ligation experiments and mass spectrometry and far-UV circular diochroism of nanostructures formed from 3 native chemical ligations.

Author Information

Corresponding Author

*Email: e.main@qmul.ac.uk

Author Contributions

ERGM & JAH conceived and designed the experiments. JAH carried out the experiments. JAH and ERGM performed the data analysis. JAH, LSI and ERGM discussed the results and guided further experiments. JAH and ERGM wrote the manuscript. All authors edited and approved the manuscript.

Acknowledgements

LSI acknowledges the support of a Senior Fellowship from the UK Medical Research Foundation. ERGM and LSI labs acknowledge support from a Leverhulme Trust project grant. JAH was supported by a QMUL Principal's studentship.

Competing financial interests

The authors declare no competing financial interests.

References

1. Mendes, A. C.; Baran, E. T.; Reis, R. L.; Azevedo, H. S., Self-assembly in nature: using the principles of nature to create complex nanobiomaterials. *Wires Nanomed Nanobi* **2013,** *5* (6), 582-612.
2. Lai, Y. T.; Tsai, K. L.; Sawaya, M. R.; Asturias, F. J.; Yeates, T. O., Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly. *J Am Chem Soc* **2013,** *135* (20), 7738-43.
3. Fletcher, J. M.; Harniman, R. L.; Barnes, F. R.; Boyle, A. L.; Collins, A.; Mantell, J.; Sharp, T. H.; Antognozzi, M.; Booth, P. J.; Linden, N.; Miles, M. J.; Sessions, R. B.; Verkade, P.; Woolfson, D. N., Self-assembling cages from coiled-coil peptide modules. *Science* **2013,** *340* (6132), 595-9.
4. Grove, T. Z.; Osuji, C. O.; Forster, J. D.; Dufresne, E. R.; Regan, L., Stimuli-Responsive Smart Gels Realized via Modular Protein Design. *J Am Chem Soc* **2010,** *132* (40), 14024-26.
5. King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; Andre, I.; Gonen, T.; Yeates, T. O.; Baker, D., Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **2012,** *336* (6085), 1171-4.
6. Sutter, M.; Greber, B.; Aussignargues, C.; Kerfeld, C. A., Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science* **2017,** *356* (6344), 1293-97.
7. Jones, S.; Thornton, J. M., Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **1996,** *93* (1), 13-20.
8. Creighton, T. E., Protein Folding. *Recherche* **1991,** *22* (230), 314-323.
9. Patterson, D. M.; Nazarova, L. A.; Prescher, J. A., Finding the Right (Bioorthogonal) Chemistry. *ACS Chem Biol* **2014,** *9* (3), 592-605.
10. Padilla, J. E.; Colovos, C.; Yeates, T. O., Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc Natl Acad Sci U S A* **2001,** *98* (5), 2217-21.
11. Grove, T. Z.; Forster, J.; Pimienta, G.; Dufresne, E.; Regan, L., A modular approach to the design of protein-based smart gels. *Biopolymers* **2012,** *97* (7), 508-17.
12. Lai, Y. T.; Cascio, D.; Yeates, T. O., Structure of a 16-nm cage designed by using protein oligomers. *Science* **2012,** *336* (6085), 1129.
13. Speltz, E. B.; Nathan, A.; Regan, L., Design of Protein-Peptide Interaction Modules for Assembling Supramolecular Structures in Vivo and in Vitro. *ACS Chem Biol* **2015,** *10* (9), 2108-15.
14. Mejias, S. H.; Sot, B.; Guantes, R.; Cortajarena, A. L., Controlled nanometric fibers of self-assembled designed protein scaffolds. *Nanoscale* **2014,** *6* (19), 10982-8.
15. Rossi, E. A.; Goldenberg, D. M.; Chang, C. H., Complex and defined biostructures with the dock-and-lock method. *Trends Pharmacol Sci* **2012,** *33* (9), 474-81.
16. Matsunaga, R.; Yanaka, S.; Nagatoishi, S.; Tsumoto, K., Hyperthin nanochains composed of self-polymerizing protein shackles. *Nature communications* **2013,** *4*, 2211.

17. Leibly, D. J.; Arbing, M. A.; Pashkov, I.; DeVore, N.; Waldo, G. S.; Terwilliger, T. C.; Yeates, T. O., A Suite of Engineered GFP Molecules for Oligomeric Scaffolding. *Structure* **2015,** *23* (9), 1754-68.
18. Baneyx, F.; Mujacic, M., Recombinant protein folding and misfolding in Escherichia coli. *Nature biotechnology* **2004,** *22* (11), 1399-408.
19. Zaher, H. S.; Green, R., Fidelity at the molecular level: lessons from protein synthesis. *Cell* **2009,** *136* (4), 746-62.
20. Briggs, A. W.; Rios, X.; Chari, R.; Yang, L.; Zhang, F.; Mali, P.; Church, G. M., Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res* **2012,** *40* (15), e117.
21. Evans, T. C., Jr.; Benner, J.; Xu, M. Q., Semisynthesis of cytotoxic proteins using a modified protein splicing element. *Protein Sci* **1998,** *7* (11), 2256-64.
22. Giessen, T. W.; Silver, P. A., A Catalytic Nanoreactor Based on in Vivo Encapsulation of Multiple Enzymes in an Engineered Protein Nanocompartment. *Chembiochem* **2016,** *17* (20), 1931-35.
23. Vila-Perello, M.; Liu, Z. H.; Shah, N. H.; Willis, J. A.; Idoyaga, J.; Muir, T. W., Streamlined Expressed Protein Ligation Using Split Inteins. *J Am Chem Soc* **2013,** *135* (1), 286-92.
24. Brune, K. D.; Buldun, C. M.; Li, Y. Y.; Taylor, I. J.; Brod, F.; Biswas, S.; Howarth, M., Dual Plug-and-Display Synthetic Assembly Using Orthogonal Reactive Proteins for Twin Antigen Immunization. *Bioconjugate chemistry* **2017,** *28* (5), 1544-1551.
25. Thrane, S.; Janitzek, C. M.; Matondo, S.; Resende, M.; Gustavsson, T.; de Jongh, W. A.; Clemmensen, S.; Roeffen, W.; van de Vegte-Bolmer, M.; van Gemert, G. J.; Sauerwein, R.; Schiller, J. T.; Nielsen, M. A.; Theander, T. G.; Salanti, A.; Sander, A. F., Bacterial superglue enables easy development of efficient virus-like particle based vaccines. *J Nanobiotechnology* **2016,** *14*, 30.
26. Liu, Z.; Zhou, H.; Wang, W.; Tan, W.; Fu, Y. X.; Zhu, M., A novel method for synthetic vaccine construction based on protein assembly. *Sci Rep* **2014,** *4*, 7266.
27. Ryadnov, M. G.; Woolfson, D. N., Self-assembled templates for polypeptide synthesis. *J Am Chem Soc* **2007,** *129* (45), 14074-81.
28. Phillips, J. J.; Millership, C.; Main, E. R. G., Fibrous Nanostructures from the Self-Assembly of Designed Repeat Protein Modules. *Angew Chem Int Edit* **2012,** *51* (52), 13132-35.
29. Chen, Q.; Sun, Q.; Molino, N. M.; Wang, S. W.; Boder, E. T.; Chen, W., Sortase A-mediated multi-functionalization of protein nanoparticles. *Chemical communications* **2015,** *51* (60), 12107-110.
30. Veggiani, G.; Nakamura, T.; Brenner, M. D.; Gayet, R. V.; Yan, J.; Robinson, C. V.; Howarth, M., Programmable polyproteams built using twin peptide superglues. *Proc Natl Acad Sci U S A* **2016,** *113* (5), 1202-7.
31. Thomas, F.; Burgess, N. C.; Thomson, A. R.; Woolfson, D. N., Controlling the Assembly of Coiled-Coil Peptide Nanotubes. *Angew Chem Int Edit* **2016,** *55* (3), 987-991.
32. Gilbert, C.; Howarth, M.; Harwood, C. R.; Ellis, T., Extracellular Self-Assembly of Functional and Tunable Protein Conjugates from Bacillus subtilis. *ACS synthetic biology* **2017,** *6* (6), 957-967.

33. Muralidharan, V.; Muir, T. W., Protein ligation: an enabling technology for the biophysical analysis of proteins. *Nature Methods* **2006,** *3* (6), 429-438.
34. Shah, N. H.; Muir, T. W., Inteins: nature's gift to protein chemists. *Chem Sci* **2014,** *5* (2), 446-461.
35. Millership, C.; Phillips, J. J.; Main, E. R. G., Ising Model Reprogramming of a Repeat Protein's Equilibrium Unfolding Pathway. *Journal of Molecular Biology* **2016,** *428* (9), 1804-17.
36. Johnson, E. C.; Kent, S. B., Insights into the mechanism and catalysis of the native chemical ligation reaction. *J Am Chem Soc* **2006,** *128* (20), 6640-6.
37. Jackrel, M. E.; Valverde, R.; Regan, L., Redesign of a protein-peptide interaction: characterization and applications. *Protein Sci* **2009,** *18* (4), 762-74.

Figure Legends

Figure 1: (a) Schematic of the recombinant fusion protein and activation method used by Main and co-workers to produce a one-pot NCL mediated polymerisation system [28]. The arrows show where Factor Xa and MESNA cause the fusion protein to be cleaved yielding an activated and polymerisation ready CTPR3ΔS protein monomer with an N-terminal cysteine & C-terminal thioester. [The structure of the CTPR3ΔS protein is based on the structure of CTPR3 (i.e. identical minus C-terminal S-helix)]. **(b)** Scheme for native chemical ligation (NCL) polymerization of activated CTPRΔS protein monomers. (1) Monomers (topologically drawn) can only react to at orthogonal N- and C-termini driving the formation a head-to-tail covalent dimer, which undergoes N-acyl rearrangement to yield a peptide bond. (2) These then dock to produce a continuous CTPR superhelix.

Figure 2: (a - c) Schematic of the designed recombinant fusion proteins, their activation & stepwise nanostructure formation. Two fusion proteins coupled with selective activation and mixing are required for successful assembly. The two fusions were termed: **(a)** "anchor" and **(b)** "linker". Each possess the protein units to be assembled (shown as protein 1 through to protein n) inserted between differing flanking affinity protein tags.
(a) The anchor & its N-terminal activation: The anchor fusion possesses an N-terminal Glutathione-S-Transferase (GST) and a C-terminal StrepTactin (Strep) affinity tag. The anchor can only be activated at its N-termini via protease cleavage of the protecting GST tag (arrow signifies cleavage site). This liberates an active N-terminal cysteine (one half of the NCL chemistry).
(b) The linkers & their C-terminal activation: The linker fusions have an N-terminal 6 X Histidine (His6), a Mxe GyrA intein (INTEIN) and a C-terminal chitin binding domain affinity tag (CBD). In contrast to the anchor, linkers can be activated at both their N- and C-termini. Initially, all linkers are only activated at their C-termini via addition of Sodium 2-mercaptoethanesulfonate (MESNA - arrow signifies cleavage site). The MESNA induces the intein to spontaneously self-cleave, producing an active C-terminal thioester (the other half of the NCL chemistry). The CBD enables removal of the cleaved intein.
(c) Schematic stepwise nanostructure formation. (NCL1) Activated anchor and linker are mixed and a spontaneous native chemical ligation takes place between the anchor's N-terminal cysteine and the linker's C-terminal thioester. The irreversible reaction produces a traceless peptide bond. The product is isolated and reactivated by protease cleavage. (NCL2) The activated fusion can then be combined with, for example, an activated linker which has a different module to undergo another iterative round of ligation. These steps can be repeated (dotted arrow) through multiple rounds of NCL to produce nanostructures composed of differing protein modules in precise spatial arrays.

Figure 3: (a - h) NCL optimsation reactions. (a & b) Reaction yield after pre-ligation incubation of either (a) N-terminal activated CTPR3ΔS anchor or (b) C-terminal activated CTPR3ΔS linker at 4 °C followed by ligation to its reaction partner. Insets: SDS-PAGE

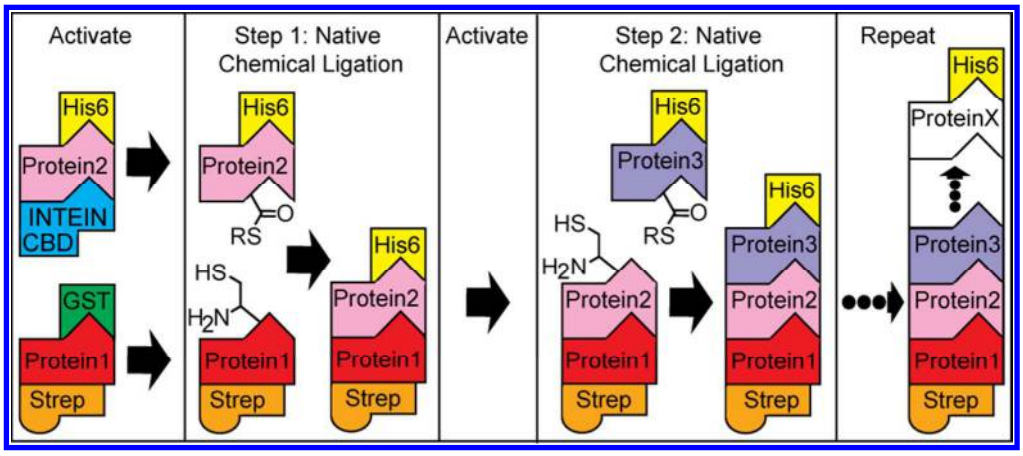
analysis of ligation yield after varying pre-ligation incubation times: Lane 1, 0 hours ligation; Lane 2, 24 hours ligation without pre-incubation; Lanes 3-7, 24 hours ligation after 24, 48, 72 and 168 hours of pre-incubation. (c -f) Quantification from SDS Page gels of the NCL reaction yield between N-terminal activated CTPR3ΔS anchor and C-terminal activated CTPR3ΔS linker over a time course of 48 hours whilst varying (c) temperature, (d) pH, (e) concentration of MESNA, (f) concentration of MPAA and after 24 hours whilst varying (g) concentration and (f) on the addition of buffer additives. Errors show 1 standard deviation. Insets in each show example SDS-PAGE gels from the time courses.

Figure 4:(a) Scheme of stepwise nanostructure formation with 3 NCL steps. Activated anchor and linker 1 were combined and a spontaneous native chemical ligation takes place between the cysteine and thioester. The irreversible reaction produces a peptide bond. The product is isolated and reactivated by TEV protease cleavage. The activated fusion was then combined with further activated linkers to undergo iterative rounds of ligation until 3 NCL reactions have occurred.

(b) SDS-PAGE analysis of 3 sequential native chemical ligations between spacer CTPR3ΔS modules. Lane 1, Protein Marker; Lane 2, activated anchor; Lane 3, activated linker; Lane 4, NCL1 0 hour; Lane 5, NCL1 24 hour; Lane 6, CTPR6ΔS product of NCL1; Lane 7, Activated of CTPR6ΔS product; Lane 8, NCL2 0 hour; Lane 9, NCL2 24 hour; Lane 10, CTPR9ΔS product of NCL2; Lane 11, Activated CTPR9ΔS product; Lane 12, NCL3 0 hour; Lane 13, NCL3 24 hour; Lane 14, CTPR12ΔS product of NCL3; Lane 15, Protein marker.

(c) A) SDS-PAGE analysis of 3 sequential native chemical ligations between spacer CTPR3ΔS anchor and binder CTPR6ΔS linker modules. Lane 1, Protein Marker; Lane 2, activated spacer CTPR3ΔS anchor; Lane 3, activated binder CTPR6ΔS linker; Lane 4, NCL1 0 hour ; Lane 5, NCL1 24 hour; Lane 6, CTPR9ΔS product of NCL1; Lane 7, activated CTPR9ΔS product; Lane 8, activated binder CTPR6ΔS linker; Lane 9, NCL2 0 hour; Lane 10, NCL2 24 hour; Lane 11, CTPR15ΔS product of NCL2; Lane 12, activated CTPR15ΔS product; Lane 13, activated binder CTPR6ΔS linker; Lane 14, NCL3 0 hour; Lane 15, NCL3 24 hour; Lane 16, CTPR21ΔS product of NCL3; Lane 17, Protein marker.

Note, masses of the CTPR proteins seem smaller than expected on the SDS PAGE gels due to “gel shifting” (they migrate faster than proteins of similar molecular weight). Mass spectrometry confirmed the molecular weights of the final products (SI Figure 5).



For Table of Contents Use Only

80x34mm (300 x 300 DPI)

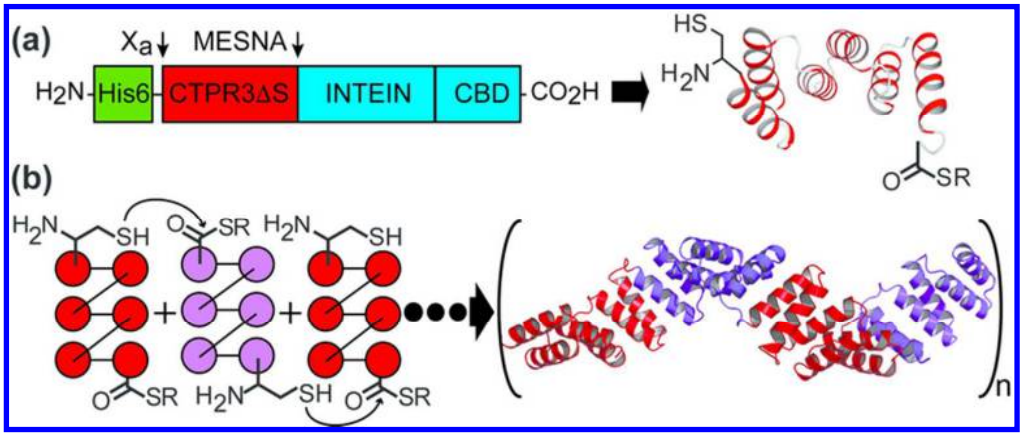


Figure 1: (a) Schematic of the recombinant fusion protein and activation method used by Main and co-workers to produce a one-pot NCL mediated polymerisation system 28. The arrows show where Factor Xa and MESNA cause the fusion protein to be cleaved yielding an activated and polymerisation ready CTPR3ΔS protein monomer with an N-terminal cysteine & C-terminal thioester. [The structure of the CTPR3ΔS protein is based on the structure of CTPR3 (i.e. identical minus C-terminal S-helix)]. (b) Scheme for native chemical ligation (NCL) polymerization of activated CTPRΔS protein monomers. (1) Monomers (topologically drawn) can only react to at orthogonal N- and C-termini driving the formation a head-to-tail covalent dimer, which undergoes N-acyl rearrangement to yield a peptide bond. (2) These then dock to produce a continuous CTPR superhelix.

35x14mm (600 x 600 DPI)

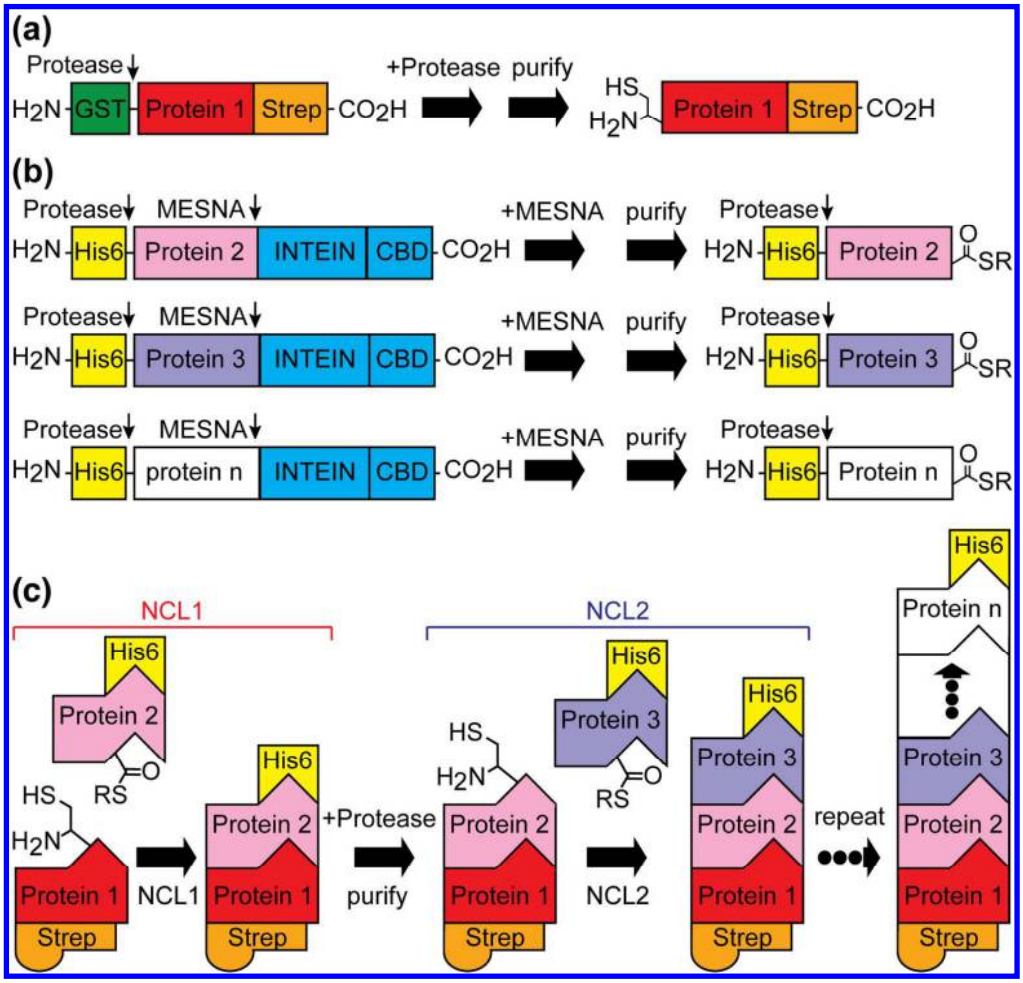
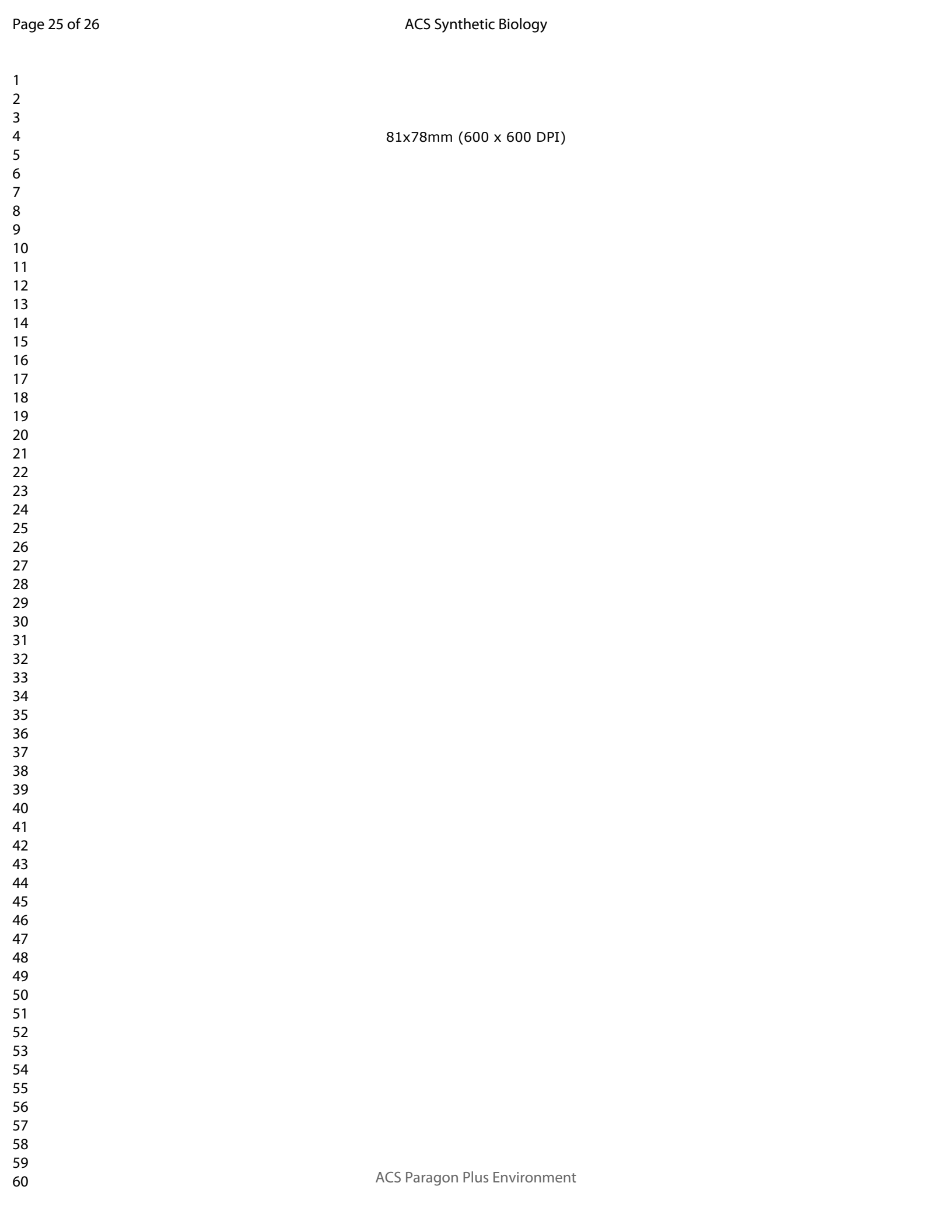


Figure 2: (a - c) Schematic of the designed recombinant fusion proteins, their activation & stepwise nanostructure formation. Two fusion proteins coupled with selective activation and mixing are required for successful assembly. The two fusions were termed: (a) "anchor" and (b) "linker". Each possess the protein units to be assembled (shown as protein 1 through to protein n) inserted between differing flanking affinity protein tags.

(a) The anchor & its N-terminal activation: The anchor fusion possesses an N-terminal Glutathione-S-Transferase (GST) and a C-terminal StrepTactin (Strep) affinity tag. The anchor can only be activated at its N-termini via protease cleavage of the protecting GST tag (arrow signifies cleavage site). This liberates an active N-terminal cysteine (one half of the NCL chemistry).

(b) The linkers & their C-terminal activation: The linker fusions have an N-terminal 6 X Histidine (His6), a Mxe GyrA intein (INTEIN) and a C-terminal chitin binding domain affinity tag (CBD). In contrast to the anchor, linkers can be activated at both their N- and C-termini. Initially, all linkers are only activated at their C-termini via addition of Sodium 2-mercaptoethanesulfonate (MESNA - arrow signifies cleavage site). The MESNA induces the intein to spontaneously self-cleave, producing an active C-terminal thioester (the other half of the NCL chemistry). The CBD enables removal of the cleaved intein.

(c) Schematic stepwise nanostructure formation. (NCL1) Activated anchor and linker are mixed and a spontaneous native chemical ligation takes place between the anchor's N-terminal cysteine and the linker's C-terminal thioester. The irreversible reaction produces a traceless peptide bond. The product is isolated and reactivated by protease cleavage. (NCL2) The activated fusion can then be combined with, for example, an activated linker which has a different module to undergo another iterative round of ligation. These steps can be repeated (dotted arrow) through multiple rounds of NCL to produce nanostructures composed of differing protein modules in precise spatial arrays.


81x78mm (600 x 600 DPI)

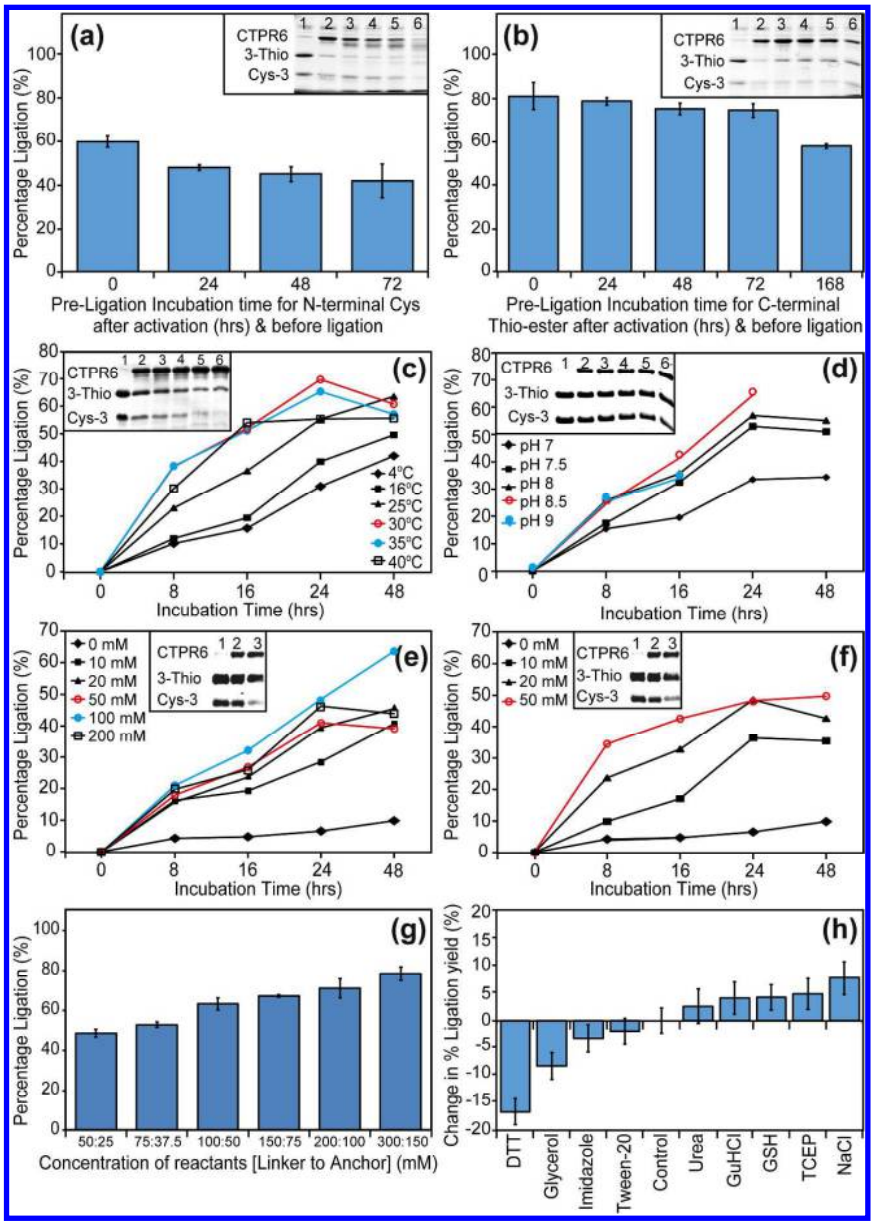


Figure 3: (a - h) NCL optimsation reactions. (a & b) Reaction yield after pre-ligation incubation of either (a) N-terminal activated CTPR3ΔS anchor or (b) C-terminal activated CTPR3ΔS linker at 4 °C followed by ligation to its reaction partner. Insets: SDS-PAGE analysis of ligation yield after varying pre-ligation incubation times: Lane 1, 0 hours ligation; Lane 2, 24 hours ligation without pre-incubation; Lanes 3-7, 24 hours ligation after 24, 48, 72 and 168 hours of pre-incubation. (c -f) Quantification from SDS Page gels of the NCL reaction yield between N-terminal activated CTPR3ΔS anchor and C-terminal activated CTPR3ΔS linker over a time course of 48 hours whilst varying (c) temperature, (d) pH, (e) concentration of MESNA, (f) concentration of MPAA and after 24 hours whilst varying (g) concentration and (f) on the addition of buffer additives. Errors show 1 standard deviation. Insets in each show example SDS-PAGE gels from the time courses.

119x167mm (600 x 600 DPI)

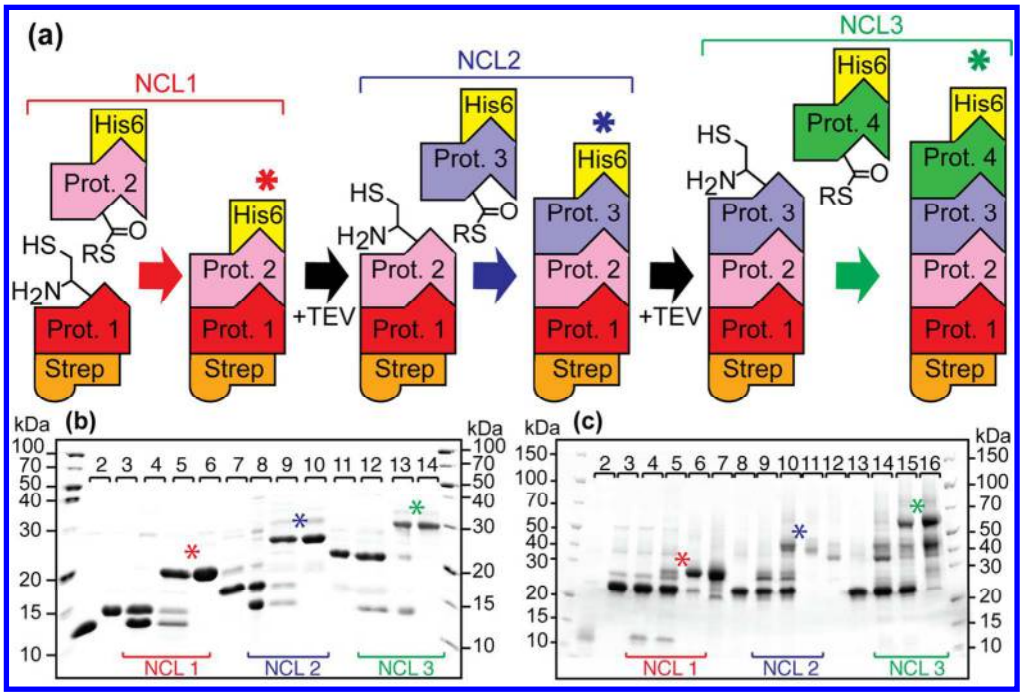


Figure 4:(a) Scheme of stepwise nanostructure formation with 3 NCL steps. Activated anchor and linker 1 were combined and a spontaneous native chemical ligation takes place between the cysteine and thioester. The irreversible reaction produces a peptide bond. The product is isolated and reactivated by TEV protease cleavage. The activated fusion was then combined with further activated linkers to undergo iterative rounds of ligation until 3 NCL reactions have occurred.
(b) SDS-PAGE analysis of 3 sequential native chemical ligations between spacer CTPR3ΔS modules. Lane 1, Protein Marker; Lane 2, activated anchor; Lane 3, activated linker; Lane 4, NCL1 0 hour; Lane 5, NCL1 24 hour; Lane 6, CTPR6ΔS product of NCL1; Lane 7, Activated of CTPR6ΔS product; Lane 8, NCL2 0 hour; Lane 9, NCL2 24 hour; Lane 10, CTPR9ΔS product of NCL2; Lane 11, Activated CTPR9ΔS product; Lane 12, NCL3 0 hour; Lane 13, NCL3 24 hour; Lane 14, CTPR12ΔS product of NCL3; Lane 15, Protein marker.
(c) A) SDS-PAGE analysis of 3 sequential native chemical ligations between spacer CTPR3ΔS anchor and binder CTPR6ΔS linker modules. Lane 1, Protein Marker; Lane 2, activated spacer CTPR3ΔS anchor; Lane 3, activated binder CTPR6ΔS linker; Lane 4, NCL1 0 hour ; Lane 5, NCL1 24 hour; Lane 6, CTPR9ΔS product of NCL1; Lane 7, activated CTPR9ΔS product; Lane 8, activated binder CTPR6ΔS linker; Lane 9, NCL2 0 hour; Lane 10, NCL2 24 hour; Lane 11, CTPR15ΔS product of NCL2; Lane 12, activated CTPR15ΔS product; Lane 13, activated binder CTPR6ΔS linker; Lane 14, NCL3 0 hour; Lane 15, NCL3 24 hour; Lane 16, CTPR21ΔS product of NCL3; Lane 17, Protein marker.
Note, masses of the CTPR proteins seem smaller than expected on the SDS PAGE gels due to "gel shifting" (they migrate faster than proteins of similar molecular weight). Mass spectrometry confirmed the molecular weights of the final products (SI Figure 5).

57x38mm (600 x 600 DPI)